Prerequisites for Advancing Automatic Speech Recognition in Breton

Alice Millour[1], Loïc Grobol[2,3], Wassim Zemouri[1,4]
Yuna Drapier[5], Mélanie Jouitteau[6]
(1) LIASD (Équipe Pastis), Université Paris 8 Vincennes Saint-Denis, France
(2) MoDyCo, Université Paris Nanterre, France
(3) Lattice, École Normale Supérieure, Montrouge, France
(4) ESI-SBA - Ecole Supérieure en Informatique 08-MAI-1945, Sidi Bel Abbès, Algérie
(5) Dastum, Rennes, France
(6) IKER, CNRS, Université de Pau et des Pays de l'Adour et
Université Bordeaux Montaigne, Bayonne, France
am@up8.edu, lgrobol@parisnanterre.fr,wassimzemouri2@gmail.com,
ydrapier@dastum.bzh, melanie.jouitteau@cnrs.fr

Abstract

We report on the extensive preliminary work of a collaborative science project aimed at developing Automatic Speech Recognition (ASR) for a minoritized European language: Breton. Hoping to help similar initiatives for other languages and communities, we present the methodology we developed for this specific ecosystem, with an estimate of the material and immaterial resources we used. Our approach is grounded in the needs and resources of the community formed by the end-users of digital development. Our multidisciplinary scientific collaboration involves linguists and speakers embedded in the academic and linguistic community, and computer scientists.

Keywords: Corpus, Transcription, ASR, less-resourced language, Dialectal Variation, Breton

1. Introduction

We report our preliminary work for a citizen science and NLP project embedded in the Breton speaking community in France called YAR (*Yezh Ar vRo*, "the language of the country" in Breton, Jouitteau et al. (2024)). This project is a collaboration between academic researchers and civil society partners: the Dastum network[1], which collects and archives recordings, an educational hub of three associations teaching Breton to adults (Roudour[2], Stumdi[3], Mervent[4]), and a translation microenterprise for the dubbing industry. It is funded by academic entities, and by an endowment fund for the digital development of Breton (Breizh Niverel[5]) in connection with local businesses and policy makers.

Requests for Breton language technologies are widespread in the community, coming from policy makers — who ask for the development of ASR technologies (Région Bretagne, 2023; Cultural Council of Brittany, 2024) — but also content creators and professionals such as teachers, dubbing and translation industry workers, developers, and fieldwork linguists. Teachers in particular have expressed the need for educational applications that include sound/transcription pairs (Ar Rouz et al., 2025). We have led preliminary interviews which report general needs for: (i) transcription of everyday languages (e.g. for text message dictation), (ii) fast indexing of raw footage for film documentaries, (iii) subtitling, and (iv) documentation — in a broad sense — of existing speech data, the scale of which prohibits manual exploration.

In this context, the aim of the YAR project is to co-design with the Breton-speaking community (i) a mobile application for geo-localized speech collection, and (ii) a web platform for collaborative transcription of speech corpora. The resources created using these collection and annotation systems will further be used to train and evaluate Automatic Speech Recognition (ASR) tools, with the hope of using them in return to make human transcription easier.

As a first step towards these solutions, we have led several exploratory actions. We report here on the following:

- A data rights awareness campaign in Brittany, in close collaboration with our partner Dastum. This includes documentation of the state of general data accessibility for the public, documentation of legal requirements, and an inventory of rights holders for archived corpora. We contacted some of the right holders to encourage them to open their datasets and supported willing ones in the process of choosing and using open licences.

- A campaign of manual transcription by linguists aimed at gathering insights on the specificities

[1] https://www.dastum.bzh
[2] https://www.roudour.bzh/
[3] https://stumdi.bzh/bzh/
[4] https://www.mervent.bzh/
[5] https://bretagnenumerique.bzh/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1464–1473
11–16 May 2026.

of transcribing Breton dialects, incrementally building transcription guidelines and ultimately providing gold standard data.

- A series of experiments on evaluating ASR solutions for Breton dialects. At this stage, we mostly focused on finding the most appropriate pre-transcription tool to support human transcription.

The practical contributions of this work are the following: (i) a state of the art of the available speech resources and ASR solutions for Breton dialects; (ii) a newly transcribed corpus of 13 h of speech: the *Korpus Treuzskrivyardurioù* (KT, Yezh Ar vRo 2025), accompanied by transcription guidelines; (iii) a quantitative and qualitative evaluation of ASR models for Breton dialects; we also report on preliminary results obtained by fine-tuning Whisper (Radford et al., 2023) on the newly available KT corpus.

2. Breton Sociolinguistic Context

A century ago, in Lower Brittany (Western France), the vast majority of people were monolingual Breton speakers. The situation is now reversed with a population estimated at 107 000 Breton speakers, scattered in a population of several millions (TMO, 2024). Rural depopulation has reconfigured the space of linguistic practice. It created diasporas in larger urban areas, which were traditionally Romance languages-speaking. All Breton speakers are now bilingual with French, and, in consequence, most of them are used to; to the digital tools available for French. It also results in widespread code-switching in natural contemporary Breton corpora, and accentuations typical of both languages coexist. Breton has vast audio and written archives, although most of them are not usable for NLP, due to the lack of proper digitization and of open rights policies.

Breton has four main traditional varieties associated to geographical areas: *Kerneveg*, *Leoneg*, *Tregerieg* (forming the said KLT group of dialects), and *Gwenedeg*. An additional fifth dialect has emerged in the KLT group during the last century: Standard Breton. It is morphosyntactically heavily influenced by Leoneg and diachronically conservative features, with some Kerneveg and Tregerieg hybridisations and some optionalities allowing for Gwenedeg variants (Jouitteau, 2020). Although Breton has known several in the last eleven centuries, schools, editors and most Breton media are currently settled on using the *peurunvan* ("completely unified") orthography, which aims to accommodate for the most salient Gwenedeg pronunciations. It is usually possible to use this standard orthography to transcribe spoken forms of all KLT+Gw dialects, but at the cost of pronunciation transparency. Some local orthographies are also still used, especially for dialectal studies or local readerships.

3. Take Stock of and Build on Available Resources

3.1. Raw Speech corpora

Extensive audio archives exist in Breton. They are often unusable for our purposes due to their unclear or restrictive licenses. Opening them requires significant work to find the rights holders, help them understand the implications of their consent (or lack thereof) and (hopefully) have them agree to use open license. It doesn't help that the technologies that can make use of their resources are very new to many of them, and are rapidly evolving.

Laws and legal directives are also evolving fast and EU and French legislations do not fully agree on intellectual property rights: EU Directive 2019/790 (European Parliament and Council of the European Union, 2019) recommends an opt-out-able exception to intellectual property rights for the purpose of data mining, but its transposition in French law reduces this practice to scientists (RF, 2021) and the official interpretation of the scope of "legitimate interest" (CNIL, 2025) was only published in July 2025.

3.1.1. Existing Raw Speech Corpora

The associative Radio Kreiz Breizh estimates its Breton archive to about 1500 h. At least four other radios archive Breton material. The archive of our partner Dastum also contains several hundreds of hours of digitized audio of spoken Breton, none of which is licensed for commercial use and about half of it has no identified owner. The archives of the Finistère department also reports 343 h in the Kerne dialect (collection *Spered ar Yezh*), whose ownership is currently debated between Dastum (which hosts the audio files) and the department itself.

These corpora include a diversity of voices in KLT dialects. The sizable time period since the 1980s ensures the presence of traditional Breton, as well as of Standard Breton. The radiophonic content would be especially valuable, because their selection criteria for interviewees is focused on diversity of content, and less so on predetermined linguistic varieties.

3.1.2. Open licensing of existing audio

We identified three key factors for opening the rights on these corpora:

Table 1: Summary of the existing open corpora and our *Korpus Treuzskrivyardurioù* (KT) in its two versions: 2025.0, which we used for our experiments in section 4, and the more recent 2025.1, which only differs in size. The figures for CV21 are those for the *validated* subset.

Dataset	Spontaneous	Dialects	Speakers	Duration (h)	Licence
CV21	no	mostly Standard	201	26.2	CC 0[†]
BSDB	yes	multidialectal	unk.	18.5	CC BY-NC-ND
KT 2025.0	yes	Kerneveg	unk.	4	CC BY-NC-SA
				3	CC BY-NC-ND
KT 2025.1	yes	Kerneveg	unk.	6	CC BY-NC-SA
				5	CC BY-NC-ND

[†] Downloading CV is subject to an agreement not to redistribute it, but the license is otherwise CC 0.

1. Providing information as to the general usability of the material for NLP,
2. Providing information about how digital tools could help each structure use their own material,
3. Providing information about what concrete steps they have to take in order to clarify the rights status of their data, and ready to use material.

We provided 1. and 2. in both public or academic conferences, and one-on-one meetings with the association's leaders and professionals handling audio corpora. Individuals provided with 1. alone showed three types of reactions. A first type was to close their data in the hope to sell it. A second type, mostly observed in language activists, was to provide it for free without the associated rights, sometimes without any option to anonymize sensitive data. A third type, mostly represented among collectors already offering data online, provided oral or written agreement via private emails, but with no public mention of the licence on their own websites. Institutions reacted rather positively to 2., but had issues finding appropriate legal counsel. The deciding factor was 3.

We have written a summary of the legal European and French situations[6], and we answer to specific questions addressed to us in a FAQ about copyrights.[7] Above all, we wrote very clear and short standard forms to copy and adapt according to very few criteria[8], and we offered help for case by case adaptation.

After six months, our data rights awareness campaign has led to concrete results and two audio archives of the Kerneveg dialect have been released under open licences: Marsel Gwilhou's (20 h, CC BY-SA) and Per Denez' (150 h, CC BY-NC-SA with the clause"if family matters are kept out"). Apart from these 20 h audio corpus, commercial use has been consistently refused so far, excluding local companies from the digital development of the language, both as users and developers.

We interpret these results as confirming that encouraging data collection with natively open source rights is an appropriate strategy for language preservation and NLP development. Policy makers would also be key here in providing support to cultural institutions. However, data rights outreach work remains essential, as it is motivated by broader ethical reasons than strictly legal ones.

3.1.3. Available Speech Corpora

Before "unlocking" the 170 h of spoken traditional Kerneveg, we had identified three main available speech corpora with adequate dialectal coverage and clear licenses, originating from CRBC[9] and transcribed part of them (see Table 1). The process provided us with a realistic view of the challenges we will encounter in the transcription platform, and will serve to populate it at its co-design stage with Breton teachers.

Enquêtes dialectologiques en vue de constituer le Nouvel Atlas Linguistique de la Basse-Bretagne (NALBB) The NALBB[10] (Le Dû, 2014) is available under CC BY-NC-ND and consists in surveys conducted with native speakers all over

[6] See: https://arbres.iker.cnrs.fr/index.php?title=Legal_guide.

[7] See: https://arbres.iker.cnrs.fr/index.php?title=FAQ_sur_le_copyright.

[8] See: https://arbres.iker.cnrs.fr/index.php?title=YAR_forms.

[9] Centre de recherche bretonne et celtique, Université de Bretagne Occidentale

[10] Dialectological surveys aiming at compiling the New Linguistic Atlas of Lower Brittany.

Lower Brittany from the 1970s to the 2010s; speakers had to provide vocabulary by translating sentences or words from French to Breton or using other forms of elicitation (naming pictures). The interviewers intervene mostly in Breton, but ask for translation of French sentences.

Atlas Linguistique des Côtes de l'Atlantique et de la Manche (ALCAM) The ALCAM[11] dataset (Le Dû, 2015), available under CC BY-NC-ND 2.5. contains twenty lexicologist surveys carried out from 1983 under the coordination of Jean Le Dû by four collectors. Vocabulary is very specialized at times.

Brezhoneg war an Dachenn (BWD) BWD[12] (Blanchard and Thomas, 2022) is available under CC BY-NC-SA: sociolinguistic surveys conducted by students with native speakers all over Lower Brittany in the 2000s and 2010s; speakers had to provide everyday and technical vocabulary by translating sentences from French to Breton. Some interviewers are not fluent in Breton, and intervene mostly in French. The speakers answer in Breton, and sometimes French.

3.2. Transcribed Corpora

To our knowledge, three legally usable transcribed corpora were available for Breton before the project started[13]. They are the following:

3.2.1. Mozilla Common Voice

The datasets collected and hosted by Mozilla Common Voice (CV[14]) project (Ardila et al., 2020), consisting of good quality recordings of short sentences read by volunteer Breton speakers and is available under a permissive license[15]. In this work we have used CV21 (Common Voice Community, 2025) to evaluate existing models.

As of CV21, the validated set has a total of 26.2 h[16] of recorded audio, and includes 201 different voices, whose gender is self- only for 65 males and 12 females. 40 % of speakers declare being below 40, and 26 % between 40 and 70. They read 7708 unique sentences, with an average duration of less than 3.2 s . As usual for *read* speech, over-articulation is pervasive for speakers of various fluencies.

The only dialectal metadata comes from an optional and rarely provided field in the profiles of speakers.

Generally speaking, several factors favour the collection of Standard Breton: the sentences to be read are mostly[17] written in the peurunvan orthography — which has a powerful priming effect for Standard Breton — ; some input sentences are sourced to the Ofis Publik ar Brezhoneg ("Public Office of the Breton Language") — which tends to be a proponent of Standard ; and some sentences pertain to typically written styles, with *e.g.* use of the simple past paradigm.

The validation process also has a bias toward Standard Breton with validators — who can be speakers of any variety of Breton — being presented with untagged writing/pronunciation pairs. Consider for example the Standard Breton sentence "*Gouzout a rit pegement e koust?*" ("Do you know how much it is?"), which appears in the validated set with the typical Gwenedeg pronunciation [gujɛd ə rət pedʒɛn e guʃt]. Validators can either ask for another sentence to judge, or answer the question "*Ha distripet mat eo bet ar frazenn?*" ("Was the sentence correctly pronounced?"), which they are left free to interpret as referring to accentuation, alignment, dialect pairing — or even the propriety of using Gwenedeg in general. The validation rationale of a given validator might also vary from sentence to sentence. The bias is not total — indeed, this particular sentence *has* been validated — but it is obvious to speakers.

Sentence validators were otherwise not particularly conservative: recent morphological creations like *paotr-splegadennoù* (litt. /man-splainations/ "mansplanations") were validated, as well as some cross-dialect translations.

3.2.2. La Banque Sonore de Dialectes Bretons (BSDB)

The BSDB[18] (Desseigne, 2018) contains 7.291 audio–transcription pairs adding up to 18.5 h, released under the CC BY-NC-ND license.

These are recordings of linguistic or cultural surveys realized by 19 volunteer collectors among speakers of traditional dialects of Breton, over a period of around ten years. A subset of eight collectors provided two transcriptions for each record-

[11] Linguistic Atlas of the Atlantic and Channel Coasts

[12] "Breton in the field"

[13] Two extra corpora have been released since then: (Guennec et al., 2022) and Roadennou (Duval-Guennoc, 2025).

[14] Unless otherwise specified, in the rest of this article, "CV" refers only to the *Breton* subsets of the Common Voice corpora.

[15] Nominally CC 0, but downloading the dataset from the project's repository requires a promise not to redistribute it.

[16] Note that the data used in the standard train/dev/test split is a significantly smaller 7.6 h due to CV's policy of sentence-level, speaker-agnostic deduplication.

[17] With only rare occurrences of widely accepted dialectal orthographic accommodations like the Gwenedeg "àr" instead of the Standard written form "war".

[18] "Breton Dialects Sound Database"

Table 2: Compared transcriptions for BSDB AD-29293-JLQ-0066 sentence “I was never using the term ‘*gourzeuiou*’”; (1) is BSDB’s “local orthography”; (2) is BSDB standardized version; (3) is a Standard Breton translation, (4) is a transcription following our guidelines, and (5) a peurunvan writing of the untranslated dialectal sentence

(1)	mé		vî	ket		lâr	james	“gourzeuiou”
	1sg		be.cond.past.3sg	neg	prog.ptc	say.INF	never	“additional.days”
(2)	me	ne	ouie	ket		lavar	james	“gourzevezhioù”
	1sg	neg	know.IMP.3sg	neg		say.INF	never	“additional.days”
(3)	me	ne	vijen	ket	o	lavar	james	“gourzevezhioù”
	1sg	ne	be.cond.past.1sg	neg	prog.ptc	say.INF	never	”additional.days”
(4)	me	'	vi’	ket	'	lâr	james	“gourzeuioù”
	1sg	ne	be.cond.past.3sg	neg	prog.ptc	say.INF	never	”additional.days”
(5)	me	ne	vije	ket	o	lavar	james	“gourzeuioù”
	1sg	ne	be.cond.past.3sg	neg	prog.ptc	say.INF	never	”additional.days”

ing: in a local orthography, and in a standardized version. The later aimed at facilitating access for readers coming from the Standard Breton dialect or “literary Breton”.

Depending on the transcribers and the linguistic distance between the recorded dialect and Standard Breton, that standardisation can be anything from a mere rewriting in the peurunvan orthography to a full translation into a linguistically different dialect. Table 2 is an illustration of the two transcriptions provided in BSDB in (1) and (2). Syntactic glosses are oursand information in lines 3, 4 and 5 are ours. In (1), the transcription is close to audio and provided in a local orthography (*lâr* “to say”). The standardized version in (2) has the words of (1) under the peurunvan orthography (*lavar* “to say”). Small functional items commonly dropped under performance were further re-established in writing, like *ne* the first part of negation in (2), whose presence in (1) is revealed by the b>v initial consonantic mutation.

From (1) to (2), both the lexical items and the syntactic structure differ. This is evidence for a translation operation taking place between two distinct linguistic varieties.[19] The partial standardisation process from (1) to (2) is useful for a Standard Breton readership because it brings closest the two linguistic varieties, but (2) is actually ungrammatical in Standard Breton and in most other Breton dialects, where the verbal morphology would here have to match the features of the 1SG subject as in (3).

While not available for mass download in its original form, the dataset has been released independently[20] with some errors, including misplaced French translations, which we have corrected for our experiments.

3.2.3. Roadennoù

The datasets collected by Duval-Guennoc (2025) add up to 5.5 h and are mostly composed of standard Breton with 9 min of some traditional Kerneveg and Tregerieg. 80 % of the dataset consists of readings of press articles, whose rights are reserved and have therefore not been used in our experiments. The 9mn 17s corpus *Komzoù Brezhoneg* is part of a much larger high-quality multidialectal aligned corpora by collector Laors Jouin and transcriber Francis Favereau. We received their written consent for CC BY-NC-SA licensing by private emails, but despite our efforts, there is still

[19] In the transcription closest to spoken data (1), the predicate of the verb “to be” is a progressive aspectual structure, headed by an unpronounced progressive particle (litt. “I would never have been at saying…”). In (2), the verb “to know” is an habituative auxiliary, with the reading “to be used to”. Its direct argument is an infinitival clause (of the type “I did not use to say…”). This standardized version of (1) is however not a translation into Standard Breton. The habitual use of the verb “to know” (2) is typical of Kerneveg, seldom documented in Leoneg and absent in Tregerieg. Furthermore, the verbal agreement in (1) is specific to this Kerne sub-dialect, because of its absence of feature matching between the 1SG pre-negation subject and its verb, which bears explicit 3SG agreement morphology. The transformation into (2) has preserved this rare dialectal agreement rule.

[20] https://huggingface.co/datasets/Bretagne/Banque_Sonore_Dialectes_Bretons

no public mention of its availability on the website Becedia[21] hosting the data.

3.2.4. Summary

The two open corpora with transcriptions add up to 42.7 h. The set of speakers presents a relatively good age covering, to the exclusion of children's voices.

In general, for the development of robust ASR there is a need for longer clips than the CV norm, and for samples with either more (meetings) or less (text messages) interactivity, and the specific prosody resulting from reading could be a confusing factor. We also need data in all KLT-Gw dialects, including Standard Breton, as well as in the varieties spoken by learners of various proficiency, heavily influenced by French, since ASR systems have to be able to deal with the particularities of learner speech — which would typically not be consider relevant by the collectors of BSDB and other long quality interviews.

Manual exploration has shown that the CV dataset does not contain usable dialectal tagging. In order to test for different training strategies, the metadata has to account for dialectal granularity, at least at the level of KLT and Standard. Given the marginalisation of Gwenedeg in the existing datasets and its internal linguistic variation, it might be wise to start with pilot studies in the KLT group including Standard Breton, while helping sister projects building resources in Gwenedeg in the meantime. Regarding transcription, the community of speakers and policy makers expect an output of ASR in the standard peurunvan orthography. However, we hypothesized as a first step that training on small corpora would yield better results by using the most phonologically transparent orthography. The transcription process thus should ideally provide both, to allow for a two-step strategy (sound>dialectal orthography>peurunvan orthography). With enough data, we will be able to challenge this hypothesis.

3.3. Transcription Campaign

One of the objectives of the YAR project is to develop a transcription platform that matches the needs of the community. To that effect, we have initiated a transcription campaign aimed not only at increasing the volume of available transcribed data, but also at observing the needs of the potential users of such platform while establishing a clear methodology for gold transcription. We need to experiment with a sizeable quantity in any given dialect, and our downstream experiments will implicate intensive collaborations in the Kerne area, so we first focus on transcriptions in the Kerneveg variety. Table 1 summarizes the content of our manually transcribed corpus.

3.3.1. Methodology

Seeking for the most efficient workflow, we have decided to sample a diversity of documents which we transcribe only partially. This provides a manually transcribed corpus with voice diversity and a dialectal spectrum inside the Kerneveg dialect, which can further be used to train better pre-transcription models to be used for the pre-annotation of the remaining part.

Our transcription standards are based on ICOR standards (Groupe ICOR, 2007), adapted to the linguistic scope and goals of our project by two Breton-speaking linguists, a phonologist and an syntactician during their transcription work. The guidelines are published online under a wiki format modifiable directly by the speaking community.[22]

Where possible, we favoured consistency with the previous transcription conventions developed for Breton ASR projects (Duval-Guennoc, 2025a) but we strictly avoid any translation or corrections. Table 2 provides in (4) an example of our transcription decisions. We mostly spell words in the Standard peurunvan orthography but spoken dialectal forms that are judged too far from standard available orthographies as they would be in (5) are written in (4) as heard. We keep track of these cases in a concordance list. The transcription include some unspoken elements in the form of punctuation and capitals, and dropped elements — which are still relevant for prosody — are marked by an apostrophe. We also include a tier dedicated to the annotation of transcription suggestions for uncertain cases — which are discussed between transcribers, but not all documents could receive a second expertise before ASR experiments.

Six months in, the guidelines were robust enough to serve as a reference for two other transcribers with unequal training in linguistics.

For this campaign, transcribers used the best — but nonetheless rather low-quality — ASR model available (Duval-Guennoc, 2025a) as a preprocessing step to segment words and utterances.[23]

4. Automatic Speech Recognition Models for Breton

Several ASR models have been pre-trained or fine-tuned on Breton datasets in the last

[21] Becedia: https://www.bcd.bzh/becedia

[22] See: https://arbres.iker.cnrs.fr/index.php?title=Guide_d%27annotation_de_transcription.

[23] Jones et al. (2025) provided an interface ensuring its accessibility.

Table 3: Model fine-tuning dataset size (in hours of transcribed audio) and performances in terms of word- and character error rates (WER and CER) (Jelinek et al., 1982) -computed using Vangberg (2025)- on BSDB (Desseigne, 2018) and on the test datasets of CV21 (Common Voice Community, 2025) and KT 2025.0 (Yezh Ar vRo, 2025). The numbers in **bold** are the best per-column (lower is better for both scores). Whisper models are used with language identification forced to Breton. To the best of our knowledge, the train datasets of the Vosk-BR and Whisper models have never been released and could intersect our test dataset (except for KT). The performances we report for these models should therefore be taken as a *ceiling* and are not directly comparable to the others. Performances on BDSB target the *Standard* version of the trancription, with the caveat that it is sometimes more than a mere transliteration of the dialectal transcription.

		BSDB		CV 21		KT 2025.0	
Model	**Size (h)**	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**
`openai/whisper-small`	n/a	1.20	1.01	1.51	1.31		
`openai/whisper-medium`	n/a	1.04	0.73	1.15	0.76		
`openai/whisper-large`	n/a	0.99	0.63	0.92	0.43		
`vosk-br`	n/a	0.75	0.47	0.35	0.19	0.85	0.63
`gweltou/wav2vec2-xls-r-300m-br`	17.5	**0.82**	**0.42**	0.43	**0.16**		
`ArzhurKoadek/whisper-small-br`	2.3	0.92	0.58	**0.39**	**0.16**	1.87	1.73
`Whisper-small-KT2025.0`	9.3	1.69	1.24	0.67	0.35	**0.60**	**0.38**

years. Vangberg and Farhat's 2023 assesment of Wav2Vec2 (Baevski et al., 2020), Whisper (Radford et al., 2023) and Coqui STT (Coqui, 2022) on CV11 (Common Voice Community, 2022) reported WERs higher than 30 %, and highlighted the effectiveness of complementing scarce speech datasets with pure text corpora[24]. More recently, Duval-Guennoc (2024) evaluated the 16 models fine-tuned for Breton available on Hugging Face Hub and reported a significant performance drop when models are evaluated on diverse corpora, using the 2 h test set of CV15 (Common Voice Community, 2023) and 32 min transcribed from a TV program (BB, unreleased). In these tests, all systems consistently perform better on CV than on BB — which was to be expected, given the nature of the corpora and the immediate availability of CV, which makes it the perfect candidate for fine-tuning. Furthermore, these results do not show a correlation between the size of the models (ranging between 8 M and 1.55 G parameters) and their performances, which were generally quite poor, with a best word error rate (WER) of 0.33 (CER 0.14).

4.1. Existing Models

We report our own evaluations on more recent corpora: the test set of CV21, BSDB and a test set consisting of 20 % of KT 2025.0, which was the version available at the time of our experiments. We compare three of the best models available (according to Duval-Guennoc (2024)) and a new model trained on KT. A summary of the results is presented in table 3.

Specifically, we compare `vosk-br` (Duval-Guennoc, 2025a), a Vosk (Alpha Cephei, 2023) model trained on an unspecified dataset, which was integrated as a pre-annotation tools in our manual transcription workflow ; `wav2vec2-xls-r-300m-br` (Duval-Guennoc, 2025b), a b(`300M` version of Wav2Vec XLS-R (Babu et al., 2022) model fine-tuned on C15 and Roadennoù ; and `ArzhurKoadek/whisper-small-br` (Koadek, 2024), a version of `openai/whisper-small` fine-tuned on CV17 (Common Voice Community, 2024).

All of these models perform significantly worse on BSDB than on CV, although possible contamination between their train corpora and CV21 could have happened.

4.2. Fine-tuning experiment

As a first experiment on adapting a model with non-CV data, we further fine-tune the Koadek (2024) model on the remaining 80 % of the 7 hours of KT 2025.0. At this stage of the project we are mostly interested in probing the influence of data diversity rather than providing a definitive model for downstream use, and since, according to Duval-Guennoc (2024), model size has a limited effect

[24] Using a raw text corpus of 8 Mwords to train the KenLM Language Model (Heafield, 2011) that supports the CTC decoders for their Wav2Vec and Coqui models resulted in better performances than the end-to-end training of Whisper.

on the performances of the models, we do not experiment with other model sizes and leave a more extensive exploration to future work.

Fine-tuning was completed in approximately 2 h over 3 epochs, using a T4 GPU 16 GB VRAM provided by Google Collab. The resulting model, `Whisper-small-KT2025.0`, outperforms every other model on the KT test set, but shows degraded performances on CV21 compared to its base model.

On the KT test set, the base model performs consistently poorly on the vocabulary, except when pronunciation is very articulated or very standard. The model we fine tuned is more efficient on the common vocabulary and on the specialized — yet frequent in the context of these interviews — vocabulary. Transcription is overall more consistent with the pronunciation, and whereas repetitions were consistently mistranscribed by the base model, we do not observe this phenomenon with the model we fine-tuned.

This highlights the need for in-depth evaluation made possible only if the evaluation corpus is diverse and documented enough.

5. Conclusions

We have presented the preliminary work for advancing ASR for Breton, preparing the ground for a collaborative science research project involving such technology. Resources critical to the work were : (i) clear guidelines for legal accuracy of corpora licensing, and proposal of standard forms to be signed. This unlocked so far 170h of recorded audio in Kerne dialect (CC BY-SA, 4.0 with an additional NC restriction for most of it), available for transcription. Natively open source data collection appear as a viable and needed alternative, (ii) clear guidelines for transcription. This allowed for normalisation of 13h transcripts, and allowed for more transcribers to join the effort. (iii) An evaluation framework which highlights the complexity of developing robust ASR systems. The Breton ASR model we fine-tune show contrasted performances.

Our next goal is to improve the evaluability of the efficiency of the models for the transcription in peurunvan orthography of all the main dialects actually spoken by Breton speakers. We will split the BSDB dataset into the main diatopic varieties it documents. We will complete our KT transcriptions with a version that reflects more our desired output with a peurunvan orthography, allowing for comparability in strategies. We have started augmenting the KT corpus with four hours of spoken contemporary Standard Breton, natively under open rights.

We have identified several needs that different entities could answer to: (i) Cultural institutions need help to establish in law what constitutes "legitimate interest" for them. Policy makers are here key to the inclusion of small local businesses in the digital development of hundred minoritized languages in France. (ii) Pre-transcription would benefit from the integration of a multilingual system which would specifically transcribe the French sections present in code-switched recordings automatically. This would require spoken language identification to enable French and Breton systems alternation. Developing such tool would benefit the other minoritized languages in a situation of bilingualism with French.

Our preliminary work has also included various exploratory contacts in the community, not reported here. They provide insight into how our initiative is perceived by the speaking community. They clarify ASR use cases, informing the accurate design for our application in development. They promote the future data collection app, and prepare its appropriation by the civil society. The data already collected and/or annotated will also serve to populate the data collection application and the transcription platform we envision. We will report on these actions later on, as an integral part of co-construction of the digital tools we wish to bring to Breton digital development.

6. Discussion and Limitations

We report on evaluations of existing models even though contamination between training and evaluation corpora for models fine-tuned on a given version of CV and evaluated on another one is highly probable — though hard to estimate. However, given the results obtained on KT, this does not affect our conclusions regarding the benefits brought by diverse transcribed corpora.

7. Bibliographical References

Alpha Cephei. 2023. Vosk speech recognition toolkit. Technical report.

David Ar Rouz, Marie Baize-Varin, and Ronan Stéphan. 2025. Préparer l'outillage d'une transmission affective du breton et de l'arabe. In Presses Universitaires de Rennes, editor, *Diversité linguistique, éducation et transmission en Bretagne*, pages 161–174.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and*

Evaluation Conference, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech 2022*, pages 2278–2282, Incheon, Korea.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

CNIL. 2025. IA : Mobiliser la base légale de l'intérêt légitime pour développer un système d'IA.

Coqui. 2022. Coqui STT.

Cultural Council of Brittany. 2024. Avis du Conseil culturel de Bretagne sur le rapport du Conseil régional « Plan de réappropriation des langues de Bretagne ».

Gweltaz Duval-Guennoc. 2024. Comparaison des modèles de reconnaissance vocale pour le breton. Technical report.

Gweltaz Duval-Guennoc. 2025a. Anaouder: Speech recognition for Breton using Vosk. Software.

Gweltaz Duval-Guennoc. 2025b. Assessing the Performance of Facebook Wav2Vec2 for Fine-Tuning a Breton ASR Model. Technical report.

European Parliament and Council of the European Union. 2019. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

Groupe ICOR. 2007. Convention ICOR.

David Guennec, Hassan Hajipoor, Gwénolé Lecorvé, Pascal Lintanf, Damien Lolive, Antoine Perquin, and Gaëlle Vidal. 2022. BreizhCorpus: a Large Breton Language Speech Corpus and its use for Text-to-Speech Synthesis. In *Odyssey Workshop 2022*, pages 263–270, Beijing, China. ISCA (International Speech Communication Association), ISCA.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

F. Jelinek, R.L. Mercer, and L.R. Bahl. 1982. Continuous speech recognition: Statistical methods. In *Classification Pattern Recognition and Reduction of Dimensionality*, volume 2 of *Handbook of Statistics*, pages 549–573. Elsevier.

Dewi Jones, Sasha Wanasky, Gweltaz Duval-Guennoc, Preben Vangberg, Leena Farhat, Mélanie Jouitteau, Loïc Grobol, and Delyth Prys. 2025. Roadmap for a Welsh/Breton linguistic AI network. Technical report, Agile Cymru.

Mélanie Jouitteau. 2020. Standard Breton, traditional dialects, and how they differ syntactically. *Journal of Celtic Linguistics*, 21(1):29–74.

Mélanie Jouitteau, Alice Millour, Jean-Yves Antoine, and Loïc Grobol. 2024. Yezh Ar Vro - The language of the country : Building the appropriation of data collection applications. In *Proceedings de LIFT2*, Orléans, France. GDR LIFT.

Arzhur Koadek. 2024. Whisper Small Br.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

RF. 2021. Ordonnance n° 2021-1518 du 24 novembre 2021 complétant la transposition de la directive 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique et modifiant les directives 96/9/CE et 2001/29/CE.

Région Bretagne. 2023. Plan de réappropriation des langues de bretagne 2024-2027.

TMO. 2024. Enquête sociolinguistique 2024 sur les langues de Bretagne.

Preben Vangberg. 2025. Universal edit distance. Python Library.

Preben Vangberg and Leena Farhat. 2023. Speech-to-text for Breton. In *Celtic Student Conference*, Glasgow, United Kingdom.

8. Language Resource References

Blanchard, Nelly and Thomas, Mannaig. 2022. *Brezhoneg war an dachenn (BWD) - Enquêtes de dialectologie et de sociolinguistique*.

Common Voice Community. 2022. *Common Voice v11*. Mozilla Foundation.

Common Voice Community. 2023. *Common Voice v15*. Mozilla Foundation.

Common Voice Community. 2024. *Common Voice v17*. Mozilla Foundation.

Common Voice Community. 2025. *Common Voice v21*. Mozilla Foundation.

Adrien Desseigne. 2018. *Banque Sonore des Dialectes Bretons*.

Gweltaz Duval-Guennoc. 2025. *Roadennoù*.

Le Dû, Jean. 2014. *Enquêtes dialectologiques en vue de constituer le Nouvel Atlas Linguistique de la Basse-Bretagne*.

Le Dû, Jean. 2015. *Atlas Linguistique des Côtes de l'Atlantique et de la Manche (ALCAM)*.

Yezh Ar vRo. 2025. *Korpus Treuzskrivyardurioù*. Zenodo, 2025.1.